Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 2): pp. 1571-1576
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.06.040

Big Data Parallel Processing With Map Reduce In Hadoop Framework

Jaswinder Singh[1*], Dr. Ashwani Sethi[2]

[1, 2]Guru Kashi University, Talwandi Sabo, Bathinda, Punjab, India.

Abstract:- Advancement of information Technolgy to effect on human life. The human life going to busy in these days within smart phone or laptop. Where people use a smart to get education or other knowdlgy in different social field, they like shoping from E-Commerce sites. More industerlists develope own E- Commerce web sites. They give more offer to customer product shopping. In this paper we will describe how to E-commerce Industries analysis million byte of data in few minutes. How to compare own bussiness with other competative bussinesses.

Keyword:- BDBA,Big Data, Hadoop,Map Reduce

Introduction:-Last few years, more users connected with the internet and data transfer and recieved in many types. The data is collected by many devices (volume), very fast (velocity), in various formats (variety) and has uncertain integrity and authenticity (veracity). This rapid growth in data is commonly dubbed as Big Data and mannerism a number of interesting engineering and technical problems. One such problem is the slow processing of all those collected amounts of data. A common data processing solution is to collect the gathered data in a single place through distributed across different devices, sometimes called Data Lake. Once the data is in a single location, distributed across different nodes technologies, like Apache Hadoop are used to perform calculations on the data.There are many challenges with this approach: storing and indexing of the data, data location optimization, filtering and verification of the data authenticity, processing of unstructured data, etc. In this position paper, we advise how to fast process big data in Apache Hadoop Framework with map-reduce.[1]

Big Data has newly become a major trend attracting both academia, research institutions, and industries, with a hidden the market of 187 billion dollars by 2019 and an increasing rate of 50% over five years. In these days pervasive andn interconnected globally, in fact, make people at the center of a continuous sensing process, where an enormous amount of data are generated and collected every minute. In particular, according to every human in the world is producing over 6 megabytes for a minute, a total of 1.7 million billion bytes of data. Many organizations have own domains their discover to develop or endure competitive, they have to deal with business cases where the high volume of data reach terabytes

and even petabytes, often with an upscale sort of datatypes to be considered. Certainly, low latency access to this huge amount of distributed data represents a competitive the advantage in the market, especially for business intelligent applications.[2]

Big Data Storage

Big statistics storage systems typically address the volume mission by using making use of distributed, shared not anything architectures. This permits addressing improved garage necessities with the aid of scaling out to new nodes imparting computational energy and garage. New machines can seamlessly be introduced to a storage cluster and the storage machine looks after dispensing the statistics between character nodes transparently.[3]

Big Data uses in Market

A special region of use instances for huge facts is the manufacturing, transportation, and logistics sector. These sectors are going throw a transformational change as part of an industry-huge trend, called "Industry 4.0", which originates in the digitization and interlinking of products, manufacturing facilities, and transportation infrastructure as a part of the developing "Internet of Things". Data usage has a profound effect in those sectors, applications of predictive analysis in preservation are main to new business models as the manufacturers of equipment are inside the nice role to provide large statistics-based protection. The evolution of cyber-physical systems for manufacturing, transportation, logistics, and other sectors brings new challenges for reproduction and planning, for monitoring, control, and interaction with machinery or facts usage packages. [4]

Analysis of Big Data

Popular facts analysis techniques, which includes MapReduce, permit the introduction of a programming model and related implementations for processing and generating massive datasets.Big information can be analyzed with the software tools generally used as an element of superior analytics disciplines consisting of predictive analytics, records mining, text analytics, and statistical evaluation. Mainstream BI software and facts visualization tools also can play a role inside the analysis process. But the semi-structured and unstructured records may not healthy well in traditional information warehouses primarily based on relational databases.[5]

Big Data Business Analytics (BDBA)

This type of analysis is used by Business management to measure the performance of their company. It also carried for calculating their position in the market and to find where they should improve their strategies. This type of analytics uses statistical methods that can be applied for a specific product or process by the company. The main motive of the company to run the business analytics is to monitor their proceed of business and to identify the disadvantages of the existing processes and highlight

meaningful data. This helps the company to know the area of improvement in its business for future growth and to handle the challenges. The business analytics plays vital role to make decisions, improves the business strategies to keep a business competitive.[6]

Hadoop Framework

Today, Hadoop is extensively used in lots of corporations as a general reason platform for disbursed garage and processing of huge facts sets on commodity pc clusters. Prominent Hadoop users encompass Yahoo, Facebook, IBM, Twitter, and Adobe.[7] Hadoop is a Big Data system structured and conveyed by Apache Foundation. It is an open-source programming utility that works in the system of PCs in corresponding to discover answers for Big Data and procedure it utilizing the MapReduce calculation. Google discharged a paper on MapReduce innovation in December 2004. This turned into the beginning of the Hadoop Processing Model. In this way, MapReduce is a programming model that permits us to perform equal and appropriated handling on colossal informational indexes [8, 9]

Parallel Processing

In MapReduce, we are dividing the work among multiple nodes and each node works with an area of the work simultaneously. So, MapReduce is predicated on Divide and Conquer paradigm which helps us to process the info using different machines. As the data is processed by multiple machines rather than one machine in parallel, the time taken to process the info gets reduced by an incredible amount as shown in the figure below (2)

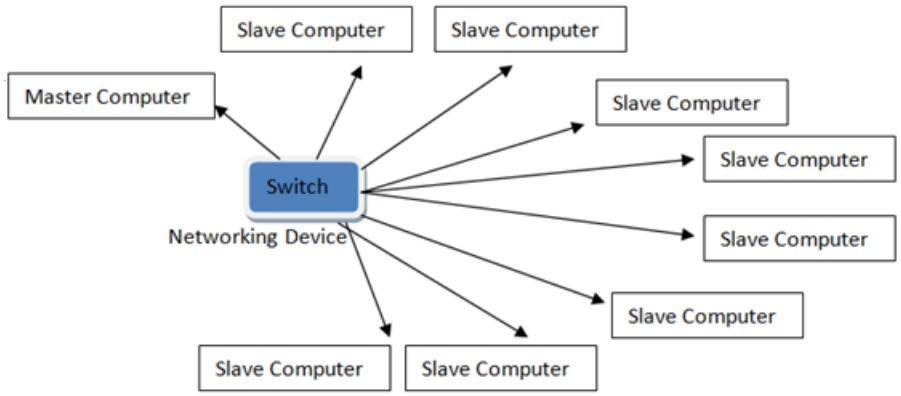


Figure: 1 Map Reduce Distributed Parallel Processing

Parallel programming developed as a way of improving performance and efficiency. In a parallel program, the processing is choppy into parts, each of which may be executed concurrently. The instructions from every part run simultaneously on different CPUs. These CPUs can exist on one machine, or they will be CPUs during a set of computers

connected via a network. Not only are parallel programs faster, they will even be wont to solve problems on large datasets using non-local resources. When you have a group of computers connected on a network, you've got a huge pool of CPUs, and you regularly have the power to read and write very large files (assuming a distributed filing system is additionally in place).[8, 9]

Hadoop Framework Experiment

Hadoop Framework developed By Google in 2004 Where Big unstructured log file process in efficient way.

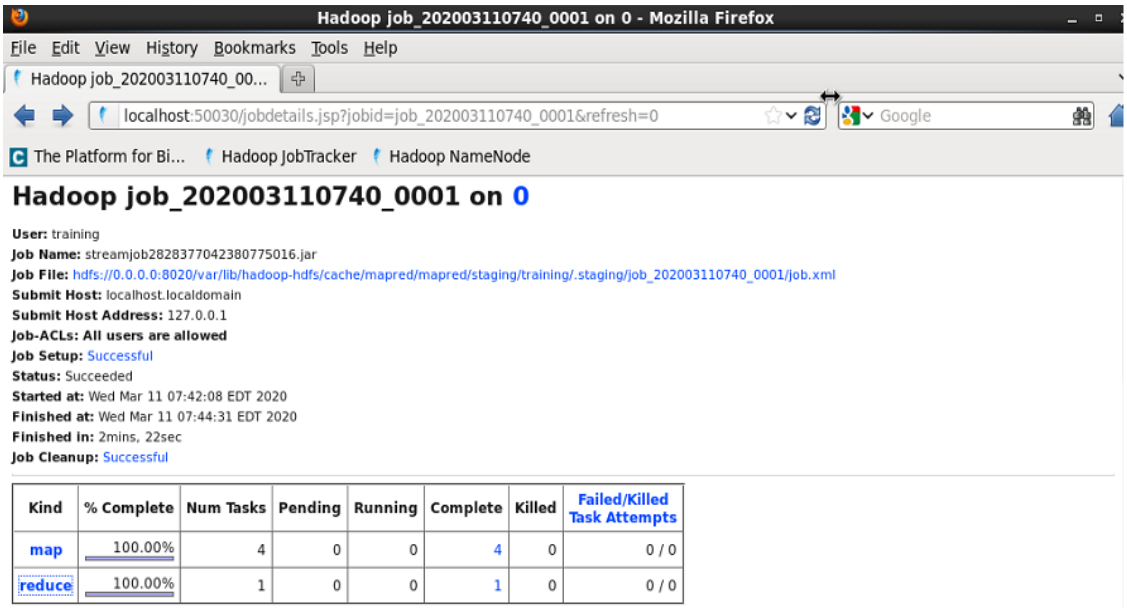


Figure : Map Reduce Task at Hadoop Framewok Machine

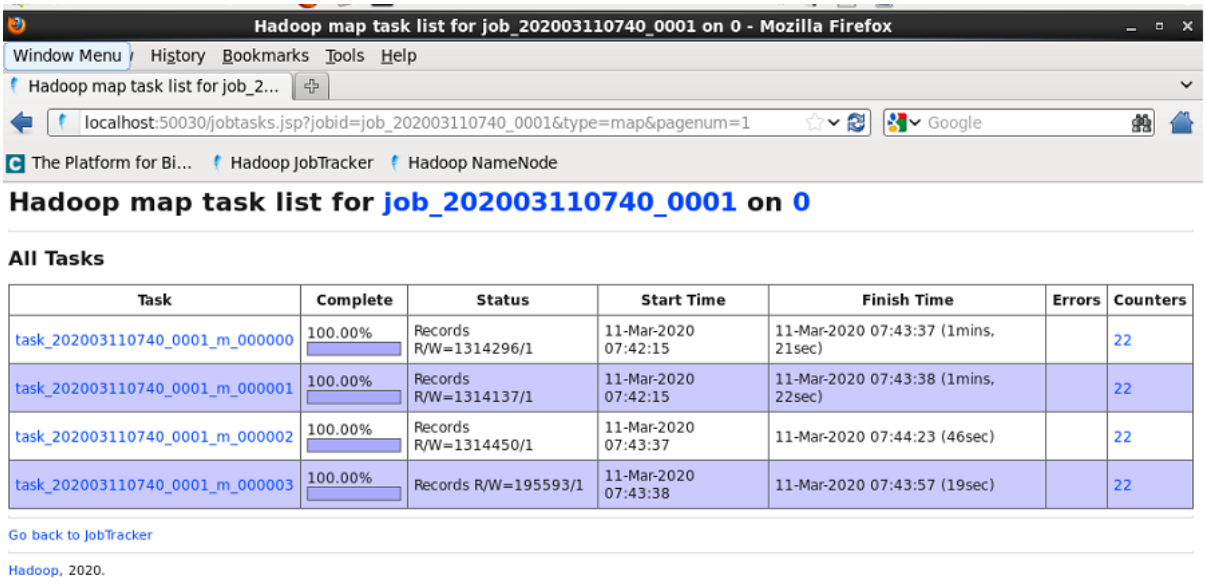


Figure: Map Task

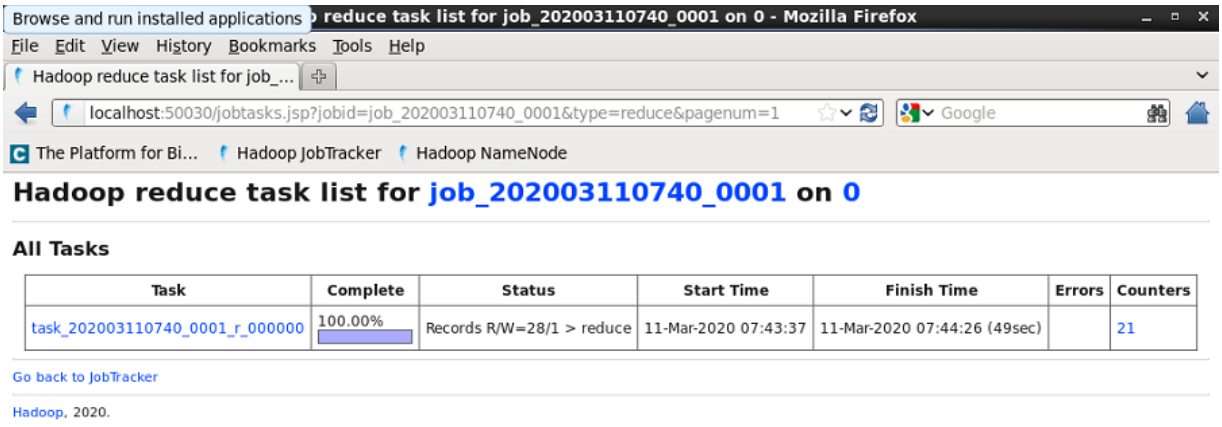
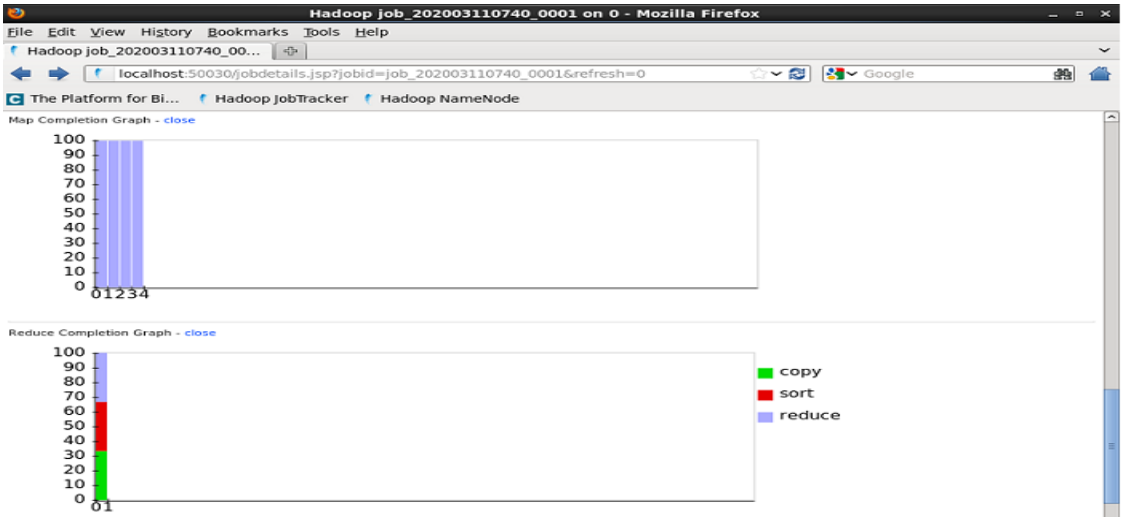


Figure: Redcue Task



Graph 1 Map Reduce Task

Conclusion

From result of above case study Our work flow based data Analytics at Hadoop Framework with Parallel Processing techinque Map Reduce. We can do Parallel work at multiple machines connected with server and run with one algorithm, As this huge data process with efficeient way. Future we can increase the perfromance of map Redcue techinque.

Reference:-

[1]. Philip Derbeko, Shlomi Dolev, Ehud Gudes ,"Concise Essence-Preserving Big Data Representation", 2016 IEEE International Conference on Big Data (Big Data),978-1-4673-9005-7/16/$31.00 ©2016 IEEE

[2]. Claudio A. Ardagna, Paolo Ceravolo, Ernesto Damiani," Big Data Analytics as-a-Service: Issues and challenges", 2016 IEEE International Conference on Big Data (Big Data), 978-1-4673-9005-7/16/$31.00 ©2016 IEEE

[3]. Martin Strohbach, Jorg Daubert, Herman Ravkin, and Mario Lischka," Big Data Storage", 2016, DOI 10.1007/978-3-319-21569-3_7

[4]. Tilman Becker," Big Data Usage", 2016, DOI 10.1007/978-3-319-21569-3_8

[5]. Imen Chebbi(B), Wadii Boulila, and Imed Riadh Farah," Big Data: Concepts, Challenges and Applications", Springer International Publishing Switzerland 2015, DOI: 10.1007/978-3-319-24306-1 62

[6]. T. Giri Babu Dr. G. Anjan Babu" A Survey on Data Science Technologies & Big Data Analytics" International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 2, February 2016

[7]. Wei Dai , Ibrahim Ibrahim, Mostafa Bassiouni," A New Replica Placement Policy for Hadoop Distributed File System", 978-1-5090-2403-2/16 $31.00 © 2016 IEEEDOI 10.1109/BigDataSecurity-HPSC-IDS.2016.30

[8] https://www.edureka.co/blog/mapreduce-tutorial

[9] Jaswinder Singh & Ashwani Sethi (2019). Big Data Store and Secure Process in Hadoop Distributed File System in Hadoop Framework. Journal of Emerging Technologies and Innovative Research, 6(6), 288-292.